2019

Predicting injury outcomes in mining industry - a machine learning approach

Anurag Desai Yedla
Iowa State University

Predicting injury outcomes in mining industry - a machine learning approach

by

Anurag Desai Yedla

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Computer Science

Program of Study Committee:
Ali Jannesari, Major Professor
Jia Liu
Steven Freeman
Gretchen Mosher

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

DEDICATION

I want to dedicate this thesis to my advisor, Dr. Jannesari Ali, for his constant support and guidance. To my parents without whom, none of this would have been possible.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I want to express my sincere thanks to my advisor Dr. Jannesari Ali, for the support, patience and guidance throughout this research. I am extremely lucky to have an advisor like him.

I would like to thank Dr. Fatemeh Davoudi Kakhki, for generously sharing her time and expertise. This research could not be completed without your generous support.

I would like to thank my committee members Dr. Jia Liu, Dr. Steven Freeman, and Dr. Gretchen Mosher, for their time and efforts through serving on this committee.

I would like to extend my wholehearted thanks to my parents, my brother, my sister and my friends. I couldn't have made this journey without you.

ABSTRACT

The mining industry plays an essential role in the US economy. Mining is known to be one of the most dangerous occupations. Even though there have been efforts to create a safer work environment for miners, there is still a significant number of accidents occurring on the mining sites. Mine operators are required to report all accidents, injuries, or illness that occurs at a mine to Mine Safety and Health Administration(MSHA). These reports contain several fixed fields entries as well as the narrative of the accident. In this study, we use machine learning models such as Decision Tree (DT), Random Forest (RF) and Deep Neural Network (DNN) to predict the outcome of the accident and the number of days the worker is going to be away from work (DAFW) using the MSHA dataset. These predictive models would be helpful for the safety experts in their efforts to create a safer work environment. Predicting days away from work would help the supervisor to plan for a temporary replacement. We compare the performance of all the models with the performance of traditional logistic regression model. We divide the study into two parts. In the first part, we use the structured data (fixed fields) and unstructured (injury narratives) separately to predict the injury outcome. We use the injury narratives because they provide more information about the accident than the fixed field entries. We also investigate the use of synthetic data augmentation technique using word embedding to tackle the data imbalance problem while predicting the injury outcome using the narratives. Our experiment results show that Random Forest with narratives as the input provides the best F1 score of 0.94. DNN has the least root mean squared error (0.62) while predicting DAFW using injury narratives as the input. The F1 score of all the underrepresented classes except one improved after the use of data augmentation technique. We use the DNN model to find the features which are most important in determining injury outcome and DAFW. We found that Nature of injury is the most important predictor of injury outcome.

CHAPTER 1. INTRODUCTION

Workplace injuries are a significant problem for many industries [1]. In 2017 alone, 2.8 million non-fatal injuries, and illnesses were reported by the private industry in the US. Nearly one-third of non-fatal workplace injuries and illnesses resulted in days away from work [2]. In 2017, 882,730 workplace injuries resulted in days away from work [2]. The number remained more or less unchanged compared to 2016. Only the manufacturing and insurance industries experienced a decline in the rate of workplace injuries. In the mining industry, there were 4,517 non-fatal lost-time injuries in the year 2015 [3]. Although the number has reduced compared to the year 2014, the injury rate has not significantly changed [3]. There is a need to improve workplace safety, especially in the mining industry, given its hazardous nature. It is crucial to analyze the injuries that have occurred previously in the industry to identify the leading causes, frame safety policies, and to predict outcomes of the injuries in the future. The cost associated with the injury (direct and indirect) is an indicator of the severity of the injury. However, the complete details of injuries in the mining industry, along with the cost, is not publicly available [4]. The Mining Safety and Health Administration (MSHA) provides datasets as part of the Open Government Initiative [5]. The dataset provides information about the injuries reported by the mine operators and contractors.

Machine learning techniques were used in many industries such as construction, railways, and agribusiness to analyze occupational injury and accidents data and build predictive models. However, the use of machine learning methods in the mining industry has been minimal. This study used supervised learning techniques to predict the outcome of the injuries in the mining industry. Days away from work (DAFW) is also an indicator of the severity of the injury. Predicting the number of days the worker is going to be away from work is useful for supervisors to manage replacements at work. This study used fully connected feedforward neural networks and Random Forest to predict DAFW due to the injuries based on the MSHA dataset.

The study is divided into two parts. In the first part of the study, fixed field entries and injury narratives in the MSHA dataset are used separately to predict the outcome of an injury. First, the fixed field entries are used to predict the injury outcome. The fixed field entries contain information such as nature of the injury, injured body part, occupation of the injured worker, job experience, activity during which the worker was injured, time of injury, source of injury, subunit of the mining site and if the mining site was a Coal or a Metal mine. Majority of the variables (features) are categorical.

There are two types of categorical variables, nominal and ordinal. Nominal variables do not have any ordering among them. Ordinal variables have an order associated with them. In MSHA data set, most of the variables are nominal variables. Categorical features have always been tricky to handle when used as input for Deep Neural Network (DNN). Some form of encoding has to be used in order to input the features to the DNN. One hot encoding is the most commonly used encoding technique. In one hot encoding, a categorical variable with n observations and d distinct values is converted into d binary variables with n observations each. For each observation, the value of the binary variable corresponding to it would be 1, and the value of other variables would be 0. However, the categorical variables we are dealing with have high cardinality. Using one hot encoding on such high cardinality variables generates many additional variables. This study used target statistics to encode the categorical variables. This form of encoding has not been used in the domain of occupational safety, where high cardinality categorical variables are very common. Then injury narratives are used to predict the injury outcome. Narratives provide additional information about the injury that is not present in the fixed field entries. The use of synthetic data augmentation using word embedding technique is investigated to tackle data imbalance problem while using narratives.

In the second part of the study, the fixed field entries and narratives are used separately to predict the days away from work. This study compares the performance of Logistic Regression, Decision Tree, Random Forest, Deep Neural Network in the predictive modeling of injury outcome and days away from work. For the models with categorical target variable, accuracy and F1 score

are used as metrics to compare the performance of the models. Accuracy is the ratio of number of correct predictions and the total number of predictions. Relying on accuracy alone can be misleading when the dataset is imbalanced. Since the dataset used in this study is imbalanced, F1 score is also used as a performance metric. F1 score is the weighted average of precision and recall. Precision is the ratio of true positives and total data points predicted as positives. Recall is the ratio of true positives and total positive data points. F1 score takes into account both false positives and false negatives. For the models with real valued target variable, root mean square error (RMSE) is used as the performance metric. The experiments conducted in this study show that Random Forest trained on injury narratives has the best F1 score of 0.94. After synthetic data augmentation, the F1 score of all the unbalanced classes increased except for one class. DNN performs better than Random Forest in predicting DAFW. DNN has a root mean square error (RMSE) of 0.62

CHAPTER 2. PREDICTING SAFETY OUTCOMES IN MINING INDUSTRY - A MACHINE LEARNING APPROACH

Modified from a manuscript submitted to IEEE International Conference on Data Mining, 2019

Anurag Yedla[a], Fatemeh Davoudi Kakhki[b], Ali Jannesari[a]

[a]Department of Computer Science, Iowa State University, Ames, IA

[b] Department of Aviation and Technology, San Jose State University, CA

2.1 Abstract

The mining industry plays an essential role in the US economy. Mining is known to be one of the most dangerous occupations. Even though there have been efforts to create a safer work environment for miners, there is still a significant number of injuries occurring at mining sites. In this study, we use machine learning models such as Decision Tree (DT), Random Forest (RF) and Deep Neural Network (DNN) to predict the outcome of the accident and the number of days the worker is going to be away from work (DAFW) using the Mine Safety And Health Administration (MSHA) dataset. These predictive models would be helpful for the safety experts in their efforts to create a safer work environment. Predicting days away from work would help the supervisor to plan for a temporary replacement. We compare the performance of all the models with the performance of traditional logistic regression model. We use structured (fixed field entries) and unstructured (text narratives of the injury) data to predict the outcome of the injury and DAFW. We also investigate the use of synthetic data augmentation technique using word embedding to tackle the data imbalance problem while predicting the injury outcome using the narratives. Our experiment results show that Random Forest with narratives as the input provides the best F1 score of 0.94. DNN has the least root mean squared error (0.62) while predicting DAFW using injury narratives as the input. The F1 score of all the underrepresented classes except one improved

after the use of data augmentation technique. We use the DNN model to find the features which are most important in determining injury outcome and DAFW. We found that Nature of injury is the most important predictor of injury outcome.

2.2 Introduction

Workplace injuries are a significant problem for many industries [1]. In 2017 alone, 2.8 million non-fatal injuries, and illnesses were reported by the private industry in the US. Nearly one-third of non-fatal workplace injuries and illnesses resulted in days away from work [2]. In 2017, 882,730 workplace injuries resulted in days away from work [2]. In the mining industry, there were 4,517 non-fatal lost-time injuries in the year 2015 [3]. Although the number has reduced compared to the year 2014, the injury rate has not significantly changed [4]. There is a need to improve workplace safety, especially in the mining industry, given its hazardous nature. It is crucial to analyze the injuries that have occurred previously in the industry to identify the leading causes, frame safety policies, and to predict outcomes of the injuries in the future. The cost associated with the injury (direct and indirect) is an indicator of the severity of the injury. However, the complete details of injuries in the mining industry, along with the cost, is not publicly available [4]. The Mining Safety and Health Administration (MSHA) provides datasets as part of the Open Government Initiative [5]. The dataset provides information about the injuries reported by the mine operators and contractors.

Machine learning techniques were used in many industries such as construction, railways, and agribusiness to analyze occupational injury and accidents data and build predictive models. However, the use of machine learning methods in the mining industry has been minimal. To the best of our knowledge, we are the first to use supervised learning techniques to predict the outcome of the injuries in the mining industry. Days away from work (DAFW) is also an indicator of the severity of the injury. Predicting the number of days the worker is going to be away from work is useful for supervisors to manage replacements at work. Our approach uses fully connected feedforward

neural networks and Random Forest to predict DAFW due to the injuries based on the MSHA dataset.

We divide the study into two parts. In the first part of the study, we use fixed field entries and injury narratives separately in the MSHA dataset to predict the outcome of an injury. First, we use the fixed field entries to predict the injury outcome. Majority of the variables (features) are categorical. Categorical features have always been tricky to handle when used as input for DNN. Some form of encoding has to be used in order to input the features to the DNN. We use categorical encoding using target statistics to encode the categorical variables. This form of encoding has not been used in the domain of occupational safety, where high cardinality categorical variables are very common. We then use injury narratives to predict the injury outcome. Narratives provide additional information about the injury that is not present in the fixed field entries. We also investigate the use of synthetic data augmentation using word embedding technique to tackle data imbalance problem while using narratives. In the second part of the study, we use the fixed field entries and narratives separately to predict the days away from work. This study compares the performance of Logistic Regression, Decision Tree, Random Forest, Deep Neural Network in the predictive modeling of injury outcome and days away from work. For the models with categorical target variable, accuracy and F1 score are used as metrics to compare the performance of the models. Accuracy is the ratio of number of correct predictions and the total number of predictions. Relying on accuracy alone can be misleading when the dataset is imbalanced. Since the dataset used in this study is imbalanced, F1 score is also used as a performance metric. F1 score is the weighted average of precision and recall. Precision is the ratio of true positives and total data points predicted as positives. Recall is the ratio of true positives and total positive data points. F1 score takes into account both false positives and false negatives. For the models with real valued target variable, root mean square error (RMSE) is used as the performance metric. Our experiments show that Random Forest trained on injury narratives has the best F1 score of 0.94. After synthetic data augmentation, the F1 score of all the unbalanced classes increased except for one class.

In brief, the rest of the paper is organized as follows: in section 2.3, related work in occupational injury analysis and the use of machine learning techniques in occupational safety have been presented. Details about the datasets are provided in section 2.4. Methods used in this study are discussed in section 2.5. Experimental analysis of the injury outcome and days away from work modeling using logistic regression, decision tree, random forest, and artificial neural network is discussed in section 2.6. Results of the experiments are provided in section 2.7. Section 2.8 presents the discussion on the results. The conclusion is presented in section 2.9. Scope of future work concludes this paper.

2.3 Review of Literature

In this section, we describe related work in the domain of occupational injuries in several industries, including mining. Kecojevica et al. (2007) performed risk analysis to characterize the injuries in mining from 1995 to 2004 [6]. It was found that there exists a severe level of risk for fatal and non-fatal days-lost injuries (NFDL), and a moderate level of risk for no days lost (NDL) injuries. Groves et al. (2007) analyzed the injuries and fatalities involving mining equipment over the period 1995 - 2004 [7]. This study found that Non-powered hand tools were the equipment category most often involved with non-fatal injuries and off-road ore haulage was the most common source of fatalities.

Several studies have used data mining and machine learning techniques to analyze injuries in other industries. Rivas et al. (2011) used data-mining techniques to analyze and predict workplace accidents [8]. The data was obtained from a questionnaire. Classification trees and Bayesian networks were used to predict if the data corresponded to an accident or an incident. Sanchez et al. (2011) used support vector machines to predict work-related accidents according to working conditions [9]. Workers were classified into two groups, those who have suffered an occupational accident in the last twelve months and those who have not. The responses of the workers to the Sixth National Survey on Working Conditions were input to the model. Tixier et al. (2016) used Random Forest and Stochastic Gradient Tree Boosting (SGBT) to predict different safety outcomes such as

type of energy involved in the accident, injury type, body part affected and injury severity [10]. Attributes describing the work environment were the inputs to the model. Davoudi et al. (2019) used Support Vector Machines, gradient boosting trees, and Naive Bayes classifier to predict injury severity in the agribusiness industry [11]. Workers compensation claims were used in the study. Severity classes were divided based on the workers' compensation monetary loss. Ramaswamy (2017) used workers' compensation claims data to characterize occupational injuries in the biofuels and commercial grain elevator industries [12].

Some studies have also used text narratives in a similar domain. Heidarysafa et al. (2018) used deep learning techniques to analyze Railway accidents' narratives [13]. In the construction industry, Goh et al. (2017) used text mining techniques to classify construction accident narratives [14]. The source of the injury was predicted from the narrative. Some studies have used machine learning and natural language processing in the field of safety in the mining industry. One such study [15] has utilized unsupervised machine learning for topic modeling mining injury narratives. Six topics were generated, and they varied by the location of the mining, degree of injury, and the year the injury occurred.

Some studies investigated injuries involving lost workdays among workers in the mining industry. Margolis et al. (2010) investigated the relationship between age, the experience of the miner to lost workdays due to injury or illness [16]. Onder (2013) used logistic regression to predict the probability of accident, resulting in higher or less than three lost workdays [17]. However, the exact number of lost workdays was not predicted. Ramaswamy (2018) used logistic regression and random forest to predict DAFW in bulk commodity handling, food manufacturing, grocery, and retail stores [12]. The study used workers compensation data provided by an insurance company. The data used in the study does not contain information about the activity, tools involved, and experience of the worker. The study also recommends the use of DNNs and random forests to model DAFW as future work.

Although machine learning techniques have been used in the domain of occupational incidents in the mining industry, predictive modeling of injury outcomes has not been done. In this study,

we use machine learning and text mining techniques to predict injury outcome and DAFW in the mining industry.

2.4 Data

MSHA accident injuries dataset was used in this study. The dataset is publicly available and was obtained from the United States Department of Labor website [18]. The dataset contains information about the accidents reported by mine operators and contractors in the USA between 2000 and 2018 [18]. After removing the records with empty values in the degree of injury column, 228,471 records remained. Each row contains 50 variables. Out of the 50 variables, 17 variables were selected. Variables containing repeated information were excluded. Variables providing information about the insurance dates which are irrelevant to the outcome of the injury and number of days away from work were removed. A new variable was added, which specifies the difference in hours between *Shift Start Time* and *Accident Time*. The categorical variables in this dataset are *Sub-unit*, *Degree of Injury*, *Mining Equipment*, *Classification*, *Accident Type*, *Occupation*, *Activity*, *Injury Source*, *Nature Injury*, *Injured Body Part*, *Degree Injury*. The continuous variables are *Job Experience*, *Hours Between Shift Start And Injury* and *Coal/Metal*. The variable *Coal/Metal* was initially a categorical variable but it was changed to a numerical variable by substituting the value "C" representing coal to 0 and "M" representing metal industry to 1. The narrative column is restricted to a certain character limit, because of which the narratives are not very long.

2.5 Methodology

For the first part of the study, *Degree Injury* is the target variable. Table 2.1 shows the description of the target classes. Both fixed field entries and injury narratives are used as input in the models separately. Similarly, in the second part of the study, *Days Lost* is the target variable. Again, we use fixed field entries and injury narratives as the input to the models. This section provides details about data pre-processing, machines learning models used in this study, word embedding, representation of narratives, and data augmentation.

Table 2.1 Codes and description for the values of Degree of Injury

Target Class Code	**Description**
Class 0	All Other Cases (Including 1^{st} Aid)
Class 1	Days Away From Work Only
Class 2	Days Restricted Activity Only
Class 3	Days Away From Work & Restricted Activity
Class 4	Fatality
Class 5	Injuries due to Natural Causes
Class 6	Injuries involving Non Employees
Class 7	No Days Away From Work, No Restricted Activity
Class 8	Occupational Illness not DEG 1-6
Class 9	Permanent Total or Permanent Partial Disability

2.5.1 Data pre-processing

Data preprocessing is the most critical step in the machine learning pipeline. Preprocessing if done well, could boost the model performance. All the rows containing empty columns were removed. All the stop words (i.e., commonly used words such as "a", "the") were removed from the injury narratives. Stemming was performed on all the words in the narratives. Stemming is the process of reducing a word to its root form, i.e., reducing the words such as "laughing", "laughed" to "laugh". Most of the variables in the fixed field entries are categorical, and some of them have high cardinality. Categorical variables with high cardinality are often challenging as input for machine learning models such as DNNs. While there are many techniques to deal with such variables, the following technique was used to encode the categorical variables.

Categorical Encoding using Target Statistics:

While one hot encoding is the most popular encoding technique, it has certain limitations. One hot encoding generates many binary variables when the cardinality of the categorical variable is high, i.e., when the categorical variable contains many distinct values. This type of encoding leads to an increase in the number of features. For high cardinality categorical features, encoding using target variable statistics can be used [19]. Let Y be a multi-valued categorical target variable, where $Y \in Y_1, Y_2, ...Y_m$. For each possible value Y_j of the target variable, a derived variable X_j is created in substitution of the original high cardinality categorical independent variable X. Each derived variable X_j will represent an estimate of $P(\frac{(Y = Y_j)}{(X = X_i)})$ using the formula shown in equation 2.1

$$S_i = \lambda(n_i)\frac{n_{iY}}{n_i} + (1 - \lambda(n_i))\frac{n_Y}{n_{TR}} \tag{2.1}$$

where, n_{TR} is the number of records, n_{iY} is the number of records belonging to class Y and n_Y is the number of records belonging to class Y. Since the sum of the probabilities is 1, creating k derived variables is redundant. So, we introduce only k - 1 derived variables and drop any one of the X_j. Generally, a function with one or more parameters is chosen as $\lambda(n)$. The parameters of $\lambda(n)$ can be adjusted depending on the data. We choose $\lambda(n)$ as a parameter function shown in equation 2.2,

$$\lambda(n) = \frac{n}{n + m} \tag{2.2}$$

where, m is a constant.

2.5.2 Word Embedding

Humans have the innate ability to understand words. But machine learning models such as DNN do not share the ability to understand words. For leveraging DNN for predictive modeling on text data, words have to be moved into a domain that the model understands. Word embedding is a good way to bridge the human understanding of words to that of a machine/model.

Word embeddings are the vector representation of words in n dimensions. When words are represented as vectors, cosine distance can be calculated between two words to check the similarity

between them. In natural language processing tasks, the performance of learning algorithms is boosted when words are represented in a vector space. In this study, Word2vec is used to transform words into vectors. Word2vec is a word embedding technique used to learn high-quality vector representation of words [20]. Word2vec trains a neural network with one hidden layer to perform a certain task. In the end, the neural network is not used for the task it is trained to perform. Instead, learning the weights of the hidden layer is the goal. There are two kinds of learning models for Word2vec. The models differ in the tasks they are trained to perform. The Skip gram model uses a target word to predict the context. *Ex. If we have the sentence "The apple is big and red" the features of apple are used to predict "the", "is", "big", "and" "red".* Whereas, the continuous bag of words (CBOW) model uses context to predict the target word. *Ex. If we have the sentence "The apple is big and red", the features of "the", "is", "big", "and", "red" are used to predict "apple".*

Skip gram model is used in this study. The neural network is trained on all narratives in the corpus. For each sentence depending on the window size, training samples are formed. *Ex. We have the sentence "The apple is big and red and the window size is two.* The training samples are generated such that each word is paired with two (window size) words in front of it separately and two words behind it separately. For the word apple the training samples generated are as follows: ("apple", "the"), ("apple", "is"), ("apple", "big"). The word apple is the input, and the word it is paired with is the target. This is done with all the words in all narratives. Since a word cannot be fed to the neural network, a one-hot vector of the word is formed. Each word is represented as a vector with 10,000 components. When encoding a word like apple as one hot vector, "1" is placed in the position corresponding to apple and 0s are placed in all other positions. The positions for all the words can be selected in any order, but they should be unique to each word. The vector with 10,000 components is fed to the input layer containing 10,000 neurons. The hidden layer contains 300 neurons, and the output layer contains 10,000 neurons. The output layer gives a single vector with 10,000 components containing probability for each word in the vocabulary to be in the context of the input word.

After the training is complete, the hidden layer is represented by 10,000 X 300 matrix. One row for each word and one column for each neuron in the hidden layer. To get the word embedding, the one-hot vector of the word (1 X 10,000 matrix) is multiplied with the hidden layer matrix (10,000 X 300 matrix). The result is a vector with 300 components which represents the word. Each word has 300 features.

2.5.3 Representation of narratives

We train the Word2vec model with the narratives in the MSHA dataset. All the narratives are divided into tokens (words) as shown in figure 2.1. Then using the trained Word2vec model, each word is represented as a vector of length 300. The vector representation of each word is multiplied with the term frequency and inverse document frequency (TF-IDF) score as shown in figure 2.2. Term frequency is the ratio of the frequency of the term in the narrative and the total number of terms in the narrative. Inverse document frequency is the logarithm of the number of the narratives in the corpus divided by the number of narratives where the specific term appears. The TD-IDF score is the product of Term frequency and Inverse document frequency. Then vector representations of the words in the narrative are added and averaged. The resulting vector is the vector representation of the narrative with 300 components. Figure 2.2 shows the process of converting narratives to vectors.

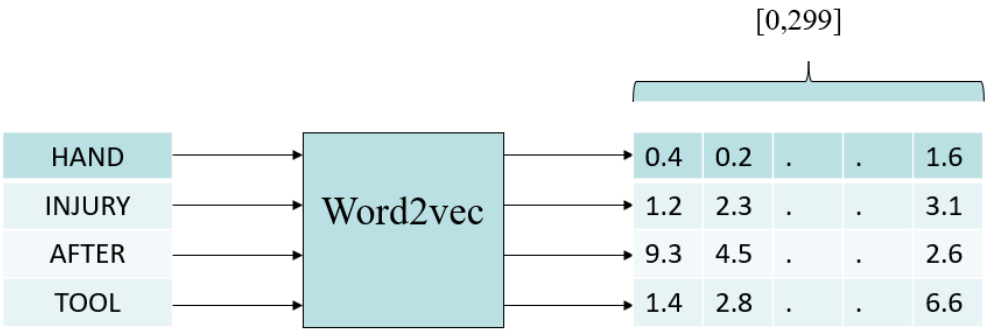

Figure 2.1 Converting each word to a vector of length 300

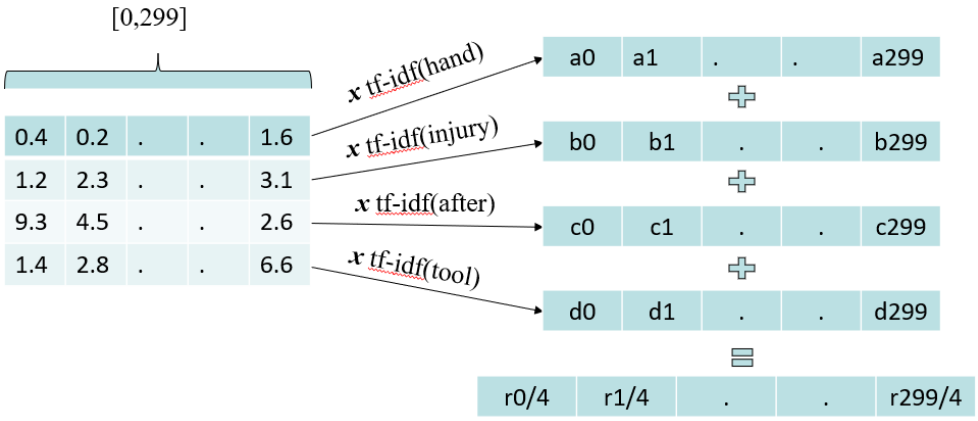

Figure 2.2 Vector representation of narratives

2.5.4 Data Augmentation

Imbalance in the target classes can often lead to the poor performance of the predictive models. The dataset used in this study is highly imbalanced. We use synthetic data augmentation to tackle the data imbalance problem. We use word embeddings to generate fake narratives [21]. First, we create ten different word2vec models, one for each target class. Then we randomly choose six words that will be replaced in each narrative in the training set. We replace each of the six words with top three closest words. The top three closest words are determined using the trained word2vec models for the respective classes. This way, we can generate 18 narratives from one narrative. We do not replace words in the narrative if it is shorter than six words.

2.5.5 Predictive Models

We investigate the performance of Logistic regression, Decision Tree, Random Forest and Artificial Neural Networks.

2.5.5.1 Logistic Regression

Logistic regression has been widely used to model the odds of an outcome in the analysis of categorical data. It was first developed by Dr. Cox in 1960 [22]. Logistic regression is used when

the target variable (dependent) is categorical. Linear regression is not suitable for classification problems as the output of the linear regression model is continuous and is not bounded. On the contrary, the output of the logistic regression model is always limited to values between 0 and 1. The logistic regression equation can be written as shown in equation 2.3

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 \tag{2.3}$$

,where the left-hand side (LHS) of the equation is the natural logarithm of the odds ratio and right-hand side (RHS) is a linear function of the independent variables. The equation can be rewritten to find the estimated probability as shown in equation 2.4.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}} \tag{2.4}$$

Logistic regression can handle any number of numerical or categorical dependent variables as shown in equation 2.5

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \tag{2.5}$$

The regression coefficients for logistic regression are calculated using maximum likelihood estimation (MLE) [23]. However, logistic regression is useful when working with a linearly separable target class.

2.5.5.2 Decision Tree

A Decision Tree is a flowchart-like structure, where each leaf node represents a class label, each non-leaf node (internal node) represents a test on a dependent variable (attribute), and each branch represents an outcome of the test. Decision trees can be converted to classification rules by tracing the path from the root node to each leaf node. The decision for a tuple X, with an unknown class label, is made by testing the attribute values of the tuple against the decision tree. Domain knowledge and parameter setting are not required to construct decision trees. These properties make decision trees a popular choice for exploratory knowledge discovery. ID3(Iterative Dichotomiser), C4.5 and CART (Classification and Regression trees) are some of the famous algorithms used to construct decision trees.

Attribute selection is an essential part of constructing the decision tree. At every level of the tree, the attributes that best partition the tuples into distinct classes are chosen. Some techniques, such as tree pruning are used to improve classification accuracy on test data or unseen data.

2.5.5.3 Random Forest

Random Forest is an ensemble model where each of the classifiers in the ensemble is a decision tree classifier [24]. Ensemble learning is the method of combining several models to make the final prediction. It helps in reducing variance, bias, and improving performance. Bagging is used to reduce the variance, and boosting is used to reduce bias. Random forest uses bagging as the ensemble method and Decision Trees as individual models. Random subsets of the dataset are created and using the subsets; Decision Trees are created. Each Decision Tree is built by selecting random attributes at each node to determine the split. All the Decision Trees participate in a majority vote, and the most popular vote is chosen as the target class or label. Random Forest reduces overfitting as it averages over the independent trees.

2.5.5.4 Deep Neural Network

The Neural Network, also known as multi-layer perceptron (MLP) is a network of artificial neurons arranged in different layers. Neural Networks are made up of three kinds of layers, an input layer, one or more hidden layers, and an output layer [24, 25]. Neural networks with more than two hidden layers are generally referred to as Deep Neural Networks (DNN). The neurons in the network are also referred to as units. Each layer is made up of neurons or units. The number of features or attributes of the tuple dictates the number of units in the input layer. Similarly, the number of units in the output layer depends on the number of class labels. We use a fully connected feedforward network where every unit in the input layer and the hidden layer is connected to every unit in the layer next to it.

2.5.6 Performance Metrics

Various evaluation metrics are available to understand the performance of the models. Performance metrics used for classification and regression tasks in this study are discussed in this section.

2.5.6.1 Accuracy

Accuracy is the ratio of the number of correct predictions and the total number of predictions. The formula for accuracy is shown in equation 2.6.

$$accuracy = \frac{true\ positive}{total\ number\ of\ samples} \tag{2.6}$$

2.5.6.2 F1 Score

Accuracy is not a good measure when the target variable classes in the dataset are unbalanced. A model which predicts the target class as the majority class for every input can achieve high accuracy score. We use F1 Score as a performance measure in this study. F1 score is the weighted average of precision and recall. Precision is the ratio of true positives and total data points predicted as positives. Recall is the ration of true positives and total positive data points. The formula for precision and recall are given in equations 2.7 and 2.8. F1 scores takes false positives and false negatives into account. It provides a more practical measure of the model's performance as it uses both precision and recall. We calculate the F1 score as shown in equation 2.9 for each target class, and their average is weighted by support (number of samples).

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{2.7}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{2.8}$$

Formula for calculating F1 score is given in equation 2.9.

$$F1score = 2 * \frac{precision * recall}{precision + recall} \tag{2.9}$$

2.5.6.3 Root mean square error (RMSE)

For regression models, RMSE is widely used as a performance metric. It represents the sample standard deviation of the differences between predicted values and observed values. RMSE is calculated, as shown in equation 2.10.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_{act} - y_{pred})^2} \tag{2.10}$$

In the equation 2.10, y_{act} represents the actual value and y_{pred} represents the predicted value.

2.6 Experiments

We divided the study into two parts, predicting the outcome of the injury and predicting days away from work. In this section, we describe the experimental setup and the architecture for the models used in each experiment.

2.6.1 Predicting outcome of the injury

Two kinds of data were used in this experiment, fixed field entries, and narratives. The target variable was the degree of injury. After removing all the target classes with entries less than 1% of the dataset, three classes remained. The dataset consisting of 127,403 entries was split into training (70%) and testing(30%) sets. Stratified random sampling was used to split the dataset.

2.6.1.1 Fixed field entries

All the variables mentioned in section 2.4 were used as the independent variables except days away from work. Logistic regression, Decision Tree, Random Forest, and DNN were used. For Decision Tree, Gini index was used as the attribute selection measure. For Random Forest, the number of Decision Trees in the forest is chosen as 30, Gini index was used as the attribute selection measure. The parameters used for DNN were as follows: four hidden layers, rectified linear units as the activation function for hidden layers, softmax as the activation function for output layer, the learning rate of 0.001 and drop out rate of 0.3.

2.6.1.2 Narratives

The input to the models was the vector representation of the injury narratives, which is computed, as shown in section 2.5. The parameters for the Decision Tree and Random Forest were the same as used for fixed field entries. DNN was trained on balanced (3 target classes), unbalanced (10 target classes) and augmented datasets. Table 2.2 shows the number of narratives added to each imbalanced class in training set using synthetic augmentation. Test dataset remained same for DNN when trained on the unbalanced and balanced dataset. The parameters used for DNN were the same as used in fixed field entries except for the number of neurons in the input layer.

Table 2.2 Number of records in each target class before and after synthetic augmentation

Target Class	Count Before Augmentation	Count After Augmentation
Class 0	669	7564
Class 1	44977	44977
Class 2	16657	16657
Class 3	10098	10098
Class 4	331	3842
Class 5	264	2785
Class 6	55	614
Class 7	27607	27607
Class 8	909	9676
Class 9	1023	12796

2.6.2 Predicting days away from work

Two kinds of data were used in this experiment, fixed field entries, and injury narratives. The target variable was the number of days lost due to the injury. The dataset consisting of 79457 records were split into training (70%) and testing (30%) using stratified random sampling. All the records with zero days lost were removed.

2.6.2.1 Fixed field entries

All the variables mentioned in section 2.4 were used as the independent variables except the degree of injury. The target variable is days away from work. Random Forest and DNN were used. For Random forest, the number of Decision Trees in the forest is chosen as 30. Mean Squared Error (MSE) was used as the function to measure the quality of a split. The parameters used for DNN were as follows: four hidden layers, rectified linear units as the activation function for hidden layers, Softplus as the activation function for output layer, the learning rate of 0.001 and drop out rate of 0.3. Softplus activation function was used as the activation function for the output layer to prevent the model from predicting negative values. MSE was used as the performance metric.

2.6.2.2 Narratives

The input to the models was the vector representation of the narratives, which is computed, as shown in section 2.5. All the parameters for Random Forest and DNN were similar to the parameters used in the fixed fields entries section of predicting days away from work except the number of neurons in the input layer for DNN. Keras and Sklearn (machine learning libraries in python) were used to build all the models.

2.7 Results

In this section, we show and compare the performance of all the models in predicting injury outcome and days away from work. The results are in two parts. In the first part, we show and compare the performance of logistic regression, Decision Tree, Random Forest, DNN (with fixed field and injury narratives as input) in predicting the injury outcome. In the second part, we show and compare the performance of logistic regression, Decision Tree, Random Forest, DNN (with fixed field and injury narratives as input) in predicting days away from work.

2.7.1 Injury Outcome

Logistic regression, Decision Tree, Random Forest, and DNN were used to predict the injury outcome. We used two kinds of inputs, fixed field entries, and injury narratives. Table 2.3 shows the overall accuracy and F1 score of the models with fixed field entries as input. All the models had decent performance except the Decision Tree. DNN had the best overall accuracy of 78%. Logistic regression and Random Forest had an accuracy of 67% and 66%. DNN was also the best model in terms of F1 score. DNN had an F1 score of 0.67. Logistic regression and Random Forest had an F1 score of 0.64 and 0.65. Logistic regression had an F1 score compared to Random Forest. Overall, DNN performed better than all other models. DT had the least accuracy (58%) and F1 score (58%).

Table 2.3 Accuracy and F1 score for all the models

Model	**F1 score**	**Accuracy**
LR	0.64	67%
DT	0.58	58%
RF	0.66	66%
DNN	0.67	78%

Table 2.4 Accuracy and F1 score for all the models

Model	**F1 score**	**Accuracy**
DNN	0.60	92%
RF	0.94	94%

Table 2.4 shows the F1 score and overall accuracy of Random Forest and DNN trained on imbalanced injury narratives. Random Forest had the highest F1 score(0.94) and accuracy (94%)among both the models. Figure 2.3 shows the confusion matrix of Random Forest trained on the injury narratives. Figure 2.4 shows the F1 score of DNN on unbalanced and balanced (using synthetic augmentation) dataset. The F1 score for all the unbalanced classes except class 5 improved after augmentation. The overall F1 score of DNN on the unbalanced dataset was 0.60. After augmentation, the overall F1 score decreased to 0.58.

True Label \ Predicted Label	Class _0	Class_1	Class_2	Class_3	Class_4	Class_5	Class_6	Class_7	Class_8	Class_9
Class _0	257	15	0	0	0	0	0	6	5	0
Class_1	0	19106	4	3	1	7	0	183	65	13
Class_2	0	543	6311	0	0	0	0	212	35	10
Class_3	1	642	9	3517	0	3	0	58	27	5
Class_4	0	2	0	0	154	0	0	0	1	0
Class_5	0	3	1	0	0	110	0	0	3	0
Class_6	0	1	0	0	0	1	18	1	0	0
Class_7	0	483	1	1	0	2	0	11272	50	9
Class_8	0	4	0	0	0	1	0	0	388	0
Class_9	0	10	0	0	0	0	0	1	0	412

Figure 2.3 Confusion matrix for Random Forest trained on injury narratives

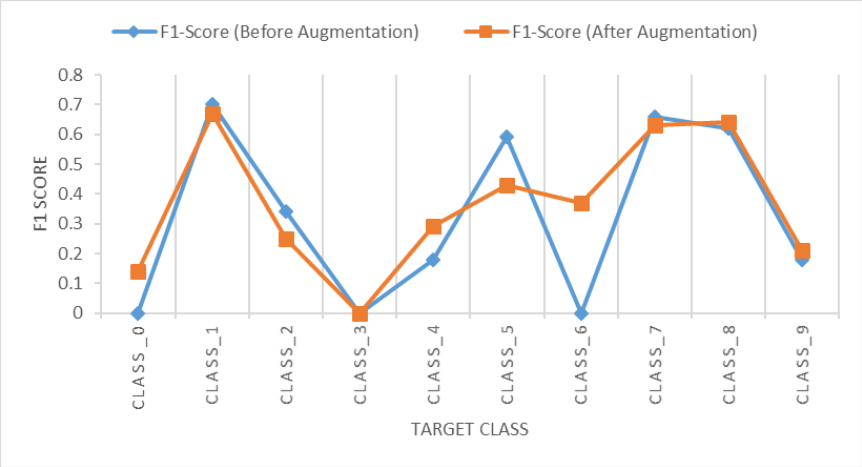

Figure 2.4 F1 score of Artificial neural network on unbalance and augmented narratives

2.7.2 Days Away from Work

a) Random Forest, DNN were used to predict DAFW. RMSE is used as the metric to compare the performance. Similar to injury outcome prediction, we used two kinds of inputs, fixed field entries, and injury narratives. The standard deviation of DAFW variable in the dataset was 75.02. Table 2.5 shows the MSE and RMSE for all the models with fixed field entries as input. DNN had the best performance compared to others. RMSE for DNN was 0.62. Random Forest had a RMSE value of 3.82. Table 2.5 shows the RMSE for DNN with injury narratives as input. Overall, DNN with fixed fields entries as input performed better than all other models.

Table 2.5 MSE and RMSE for all the models

Model	Input	MSE	RMSE
RF	Fixed Field Entries	14.65	3.82
	Injury Narratives	1502.61	38.76
DNN	Fixed Field Entries	0.38	0.62
	Injury Narratives	5944.74	77.10

2.8 Discussion

Prediction of injury outcome was accomplished using supervised machine learning techniques. The experiments done in this study show that Random Forest trained on the vector representation of injury narratives performed better than all other models. The high accuracy and an F1 score of Random Forest even when the classes are unbalanced shows the effectiveness of ensemble learning methods. Using the information in the narratives that are not present in the fixed field entries could be one of the reasons for superior performance. DNN performed relatively better than other models when the input was fixed field entries. However, underrepresented classes were removed from the dataset when fixed field entries were used.

Due to the unstructured nature of text narratives, it is not possible to identify the features which are most helpful to predict the target class. So we use the DNN trained on fixed fields to analyze the feature importance. We remove one independent variable (feature) from the dataset at a time and then train the DNN. We compute the difference between the overall F1 score of the DNN trained on the complete dataset and the DNN trained on the dataset with one missing feature. We take this difference as the feature importance. Table 2.6 lists the features in the descending order of the feature importance. According to Table 2.6, Nature of Injury is the most influential feature in the dataset. Since our focus is injuries causing lost days of work, we analyze the Nature of Injury variable for the injuries resulting in Days Away from work. The highest number of nature of injuries resulting in DAFW were sprain, disc rupture, fracture, cut, laceration, bruise. The second most influential variable was Injured body part. The injuries to back, spine, s-cord, tailbone are among

the highest to result in DAFW. Occupation is also one of the essential features. An injury to the workers having the following Occupations Maintenance man, Mechanic, Repair/Serviceman, Boilermaker, Fueler, Tire tech, and Field Service tech has the highest probability to result in DAFW class.

Although the model with the best performance cannot be used to analyze feature importance, it can certainly help to answer questions such as "if this kind of injury were to happen, what would it result in?", "What if a different body part was injured rather than the body part mentioned in the narrative?". Answers to such questions would help safety managers to plan for accidents that could occur in the future.

Table 2.6 Dependent variables and their description in descending order of their importance

Feature	**Description**
Nature of Injury	Identifies the injury in terms of its principal physical characteristics.
Injured body part	Identifies the body part affected by an injury.
Occupation	Occupation of the accident victim's regular job title.
Coal or Metal	Identifies if the accident occurred at a Coal or Metal/Non-Metal mine.
Job Experience	Experience in the job title of the person affected calculated in the decimal year.
Hours	Time difference between accident time and shift begin time in hours.
Injury Source	Identifies the object, substances, exposure or bodily motion which directly produced or inflicted the injury.
Classification	Identifies the circumstances which contributed most directly to the resulting accident.
Activity	Specific activity the accident victim was performing at the time of the incident.
Accident type	Identifies the event which directly resulted in the injury/accident.
Sub-unit	The Sub-unit of the mining site where the accident occurred.

The data augmentation using word embedding increased the F1 score of DNN for unbalanced classes except for one class. But the overall F1 score of the model decreased from 0.60 to 0.58. One of the reasons for the decrease in the overall accuracy could be the way the words to be replaced are chosen. Since they are chosen randomly, the target class of the fake narrative could have changed from the target class of the original narrative. Having longer narratives would have helped in generating more accurate synthetic narratives generation.

DNN has the best performance in predicting DAFW. Accurately predicting DAFW could help the supervisors managing the workforce to plan for replacements when an injury occurs. DAFW is also an indicator of the severity of the injury. These models could be used to predict the outcome and DAFW rather than waiting for several days to find out the outcome. These models are not a replacement to an expert in safety; instead, they are tools to help safety experts to act proactively to reduce workplace injuries.

2.9 Conclusion

We explore a new research problem of predicting the outcome of the injury and the number of days away from work in the mining industry. We use logistic regression, Decision Tree, Random Forest, and DNN to predict the outcome of the injury. We used structured (fixed field) and unstructured (text narratives) data separately to build the models. We used target based statistics to encode categorical variables. This technique helped to tackle the problem of high cardinality categorical variables. Random Forest trained on injury narratives performed better than all the models. The high predictive power of the model trained on narratives, suggests that the narratives contain additional important information compared to the fixed field entries. The synthetic data augmentation with word embedding is used to tackle the data imbalance problem. This technique improved the F1 score of DNN for all the unbalanced classes except for one class. However, the overall accuracy and F1 score of the model decreased after augmentation. There is a lot of unstructured data available compared to the structured data, and the results of this study show

that using unstructured data such as text narratives could be useful in understanding the injuries better. This study shows that there is a potential for using natural language processing (NLP) and text analytics in this field. Most influential features for the prediction of outcome of the injury were listed with the help of the random forest model trained in this study. These features and their values contributing to injuries resulting in DAFW can be analyzed.

We used Random Forest, and DNN to predict the days away from work. DNN trained on fixed field entries was the best performing model with an RMSE of 0.62. Different characteristics of an injury could be input to the model, and the resulting DAFW could be analyzed. Also, the staffing manager can plan for replacement beforehand by predicting the DAFW.

Some limitations are noted for this study. The dataset used in this study, which is maintained by MSHA, could contain errors. The accidents that could have happened but did not happen might not have been recorded. The scope of the analysis was restricted to the available data. The best performing model for predicting the outcome of the injury (Random forest with narratives as training data) cannot be used to find out the most important features since the features for this model are real numbers. During data augmentation, the words to be replaced are randomly chosen. The random selection and replacement of words in the narratives might affect the target variable. The assumption that even after replacement of the words, the target class will remain the same may not be valid in all cases.

2.10 Future Work

Convolutional Neural Networks and Recurrent Neural Networks are widely applied to text classification problems. Use of such deep learning models could be investigated in occupational safety in the Mining industry. Other variables such as weather information (temperature, humidity), age of the worker could be used with these models. The narratives used in this study are concise. Extended reports on injuries can be used in future studies.

CHAPTER 3. CONCLUSION AND FUTURE WORK

3.1 Conclusion

We explore a new research problem of predicting the outcome of the injury and the number of days away from work in the mining industry. We use logistic regression, Decision Tree, Random Forest, and DNN to predict the outcome of the injury. We used structured (fixed field) and unstructured (text narratives) data separately to build the models. We used target based statistics to encode categorical variables. This technique helped to tackle the problem of high cardinality categorical variables. Random Forest trained on injury narratives performed better than all the models. The high predictive power of the model trained on narratives, suggests that the narratives contain additional important information compared to the fixed field entries. The synthetic data augmentation with word embedding is used to tackle the data imbalance problem. This technique improved the F1 score of DNN for all the unbalanced classes except one class. However, the overall accuracy and F1 score of the model decreased after augmentation. There is a lot of unstructured data available compared to the structured data, and the results of this study show that using unstructured data such as text narratives could be useful in understanding the injuries better. This study shows that there is a potential for using natural language processing (NLP) and text analytics in this field. Most influential features for the prediction of outcome of the injury were listed with the help of the random forest model trained in this study. These features and their values contributing to injuries resulting in DAFW can be analyzed.

We used Random Forest, and DNN to predict the days away from work. DNN trained on fixed field entries was the best performing model with an RMSE of 0.62. Different characteristics of an injury could be input to the model, and the resulting DAFW could be analyzed. Also, the staffing manager can plan for replacement beforehand by predicting the DAFW.

Some limitations are noted for this study. The dataset used in this study, which is maintained by MSHA, could contain errors. The accidents that could have happened but did not happen might not have been recorded. The scope of the analysis was restricted to the available data. The best performing model for predicting the outcome of the injury (Random forest with narratives as training data) cannot be used to find out the most important features since the features for this model are real numbers. During data augmentation, the words to be replaced are randomly chosen. The random selection and replacement of words in the narratives might affect the target variable. The assumption that even after replacement of the words, the target class will remain the same may not be valid in all cases.

3.2 Future Work

The following are the recommendations for future work:

- Convolutional Neural Networks and Recurrent Neural Networks are widely applied to text classification problems. Use of such deep learning models could be investigated in occupational safety in the Mining industry.

- Other variables such as weather information (temperature, humidity), age and gender of the worker could be augmented to the dataset and model the injury outcome and DAFW.

- Generative adversarial networks (GANs) have been widely used to produce artificial data which are very similar to the original data. Use of GAN in generating narratives could be experimented in future studies.

REFERENCES

[1] N. S. Hongwei Hsiao, "Occupational Injury Prevention Research in NIOSH," Safety and Health at Work., vol. 1, no. 2, pp. 107-111, 2010.

[2] Bureau of Labor Statistics, "NEWS RELEASE," 8 November 2018. [Online]. Available: https://www.bls.gov [Accessed 15 April 2019]

[3] National Institure of Occupation Safety and Health, "Number and rate of mining nonfatal lost time injuries," [Online]. Available: https://www.cdc.gov/niosh/mining

[4] J. Heberger, "Demonstrating the financial impact of mining injuries with the "Safety Pays in Mining" web application.," Mining Engineering, vol. 70, no. 12, pp. 37-43, 2018.

[5] U. S. Government, "U.S. Open Government Initiatives," [Online]. Available: https://open.usa.gov/. [Accessed 15 April 2019].

[6] V. Kecojevica, D. Komljenovic, W. Groves and M. Radomsky, "An analysis of equipment-related fatal accidents in U.S. mining operations: 1995-2005," Safety Science, vol. 45, no. 8, pp. 864-874, 2007.

[7] W. Groves, V. Kecojevic and D. Komljenovic, "Analysis of fatalities and injuries involving mining equipment," Journal of Safety Research, vol. 38, no. 4, pp. 461-470, 2007.

[8] T. Rivas, M. Paz, J.E. Martin, J.M. Matias, J.F.Garcia and J.Taboada, "Explaining and predicting workplace accidents using data-mining techniques," Reliability Engineering and System Safety, vol. 96, pp. 739-747, 2011.

[9] A. Sanchez, P. Fernandez, F. Lasheras, F.J.de Cos Juez and P.J.Garca Nieto, "Prediction of work-related accidents according to working conditions," Applied Mathematics and Computation, vol. 218, no. 7, pp. 3539-3552, 2011.

[10] A. J.P. Tixier, M. R. Hallowell, B. Rajagopalan and D. Bowman, "Application of machine learning to construction injury prediction," Automation in Construction, vol. 69, pp. 102-114, 2016.

[11] F. Davoudi. Kakhki, S. A. Freeman and G. A. Mosher, "Evaluating machine learning performance in predicting injury severity in agribusiness industries," Safety Science, vol. 117, pp. 257-262, 2019.

[12] S. Ramaswamy, "Analysis of workers' compensation claims data for improving safety outcomes in agribusiness industries", Ph.D, Dept of Industrial and Agricultural Technology Iowa State University, 2017.

[13] M. Heidarysafa, K. Kowsari, L. E. Barnes and D. E. Brown, "Analysis of Railway Accidents' Narratives Using Deep Learning," in IEEE International Conference on Machine Learning and Applications, 2018.

[14] Y. Goh and . C. Ubeynarayana, "Construction accident narrative classification: An evaluation of text mining techniques," Accident Analysis and Prevention, vol. 108, pp. 122-130, 2017.

[15] D. Passmore, C. Chae, Y. Kustikova, R. Baker and J. H. Yim, "An exploration of text mining of narrative reports of injury incidents to assess risk," in VI International Scientific Conference "Integration, Partnership and Innovation in Construction Science and Education" (IPICSE-2018), 2018.

[16] K. A. Margolis, "Underground coal mining injury: A look at how age and experience relate to days lost from work following an injury," Safety Science, vol. 48, no. 4, pp. 417-421, 2010.

[17] S. Onder,"Evaluation of occupational injuries with lost days among opencast coal mine workers through logistic regression models," Safety Science, vol. 59, pp. 86-92, 2013.

[18] United States Department of Labor, "Mine Safety and Health Administration," [Online]. Available: https://arlweb.msha.gov [Accessed 6th February 2019].

[19] D. Micci-Barreca, A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems, ACM SIGKDD Explorations Newsletter Homepage archive, Volume 3, Issue 1, July 2001, Pages 27-32, 2001.

[20] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," CoRR, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781.

[21] L. Qi, R. Li, J. Wong, W. Tavanapong and D. A. M. Peterson, "Social Media in State Politics: Mining Policy Agendas Topics," in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2017.

[22] J. R. Wilson and K. A. Lorenz, Modeling Binary Correlated Responses using SAS, SPSS and R, Springer International Publishing, 2015.

[23] S. A. Czepiel, "Maximum Likelihood Estimation of Logistic," [Online]. Available: https://czep.net/stat/mlelr.pdf. [Accessed 11 February 2019].

[24] J. Han, Data mining Concepts and Techniques, Morgan Kaufman, 2011.

[25] F. Rosenblatt, "The Perceptron : A Probabilistic model for information storage and organization in the brain," Psychological Review, vol. 65, no. 6, pp. 386-408, 1958.

APPENDIX. ADDITIONAL MATERIAL

Table .1 Description of the features and the dependent variable in the MSHA dataset

Feature	**Description**
Hours between shift-start time and injury time	Date the accident/injury/illness occurred
Degree Injury	Description of the degree of injury/illness to the individual
Subunit	Description of the subunit where the accident/injury/illness occurred
Mining Equipment	Description of the type of mining equipment involved in the accident
Shift Begin Time	Time the shift started during which the incident occurred
Classification	Description of the accident classification that identifies the circumstances which contributed most directly to the resulting accident.
Accident Type	The accident type identifies the event which directly resulted in the reported injury/accident.
Job Experience	Experience in the job title of the person affected calculated in the decimal year. The calculation uses both the years and months of experience.
Occupation	Occupation of the accident victim's regular job title.
Activity	Specific activity the accident victim was performing at the time of the incident.
Injury Source	The source of injury identifies the object, substances, exposure or bodily motion which directly produced or inflicted the injury.
Nature Injury	The nature of injury identifies the injury in terms of its principal physical characteristics.
Injured Body Part	Identifies the part of the body affected by an injury.
Days Lost	Actual days lost from work due to the injury/illness.
Coal Metal Industry	Identifies if the accident occurred at a Coal or Metal/Non-Metal mine.